Product Sales Analysis

Amitsingh Pardeshi, Abdullah Ejaz

December 10, 2017

Project Report

1 Repository

Please find the link to the git repository- Project Link

2 Abstract

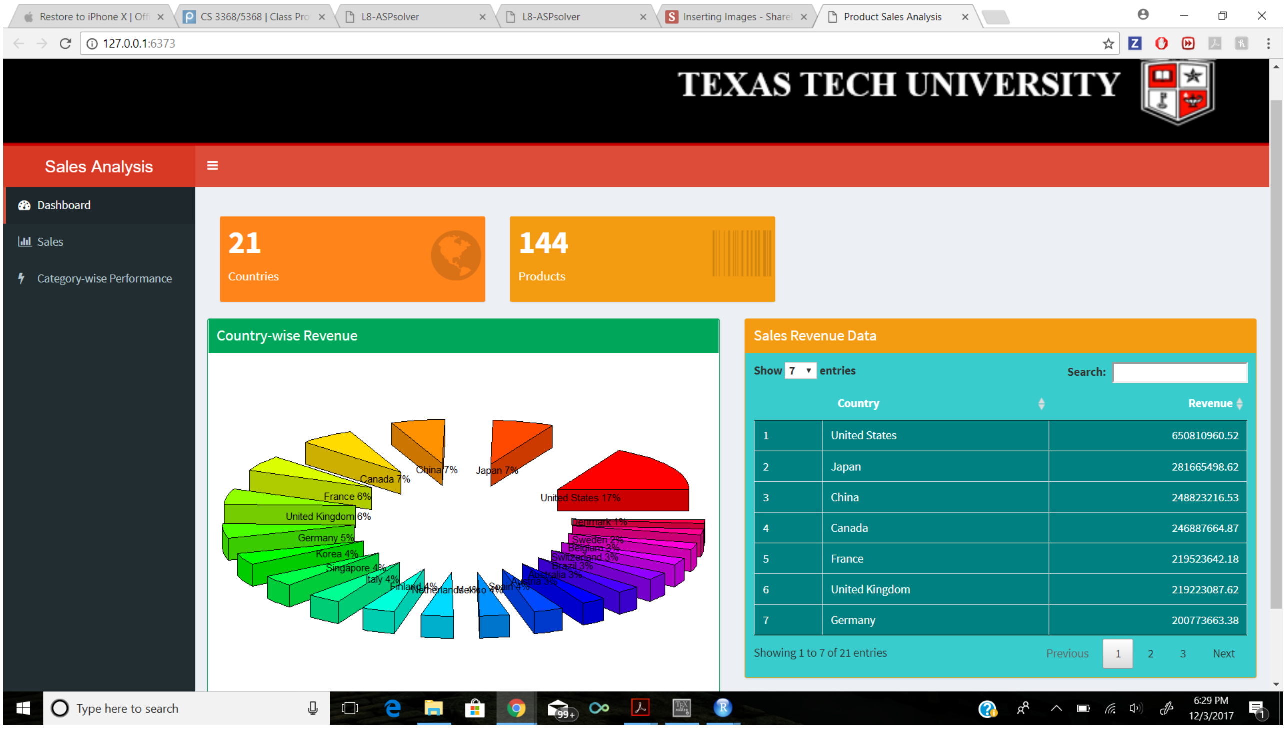
Designing an interactive web application to interpret the sales data of 21 countries with all their retailers and their different product lines. The analysis will provide us with the results that will help us identify the least performing countries and their respective retailers. This analysis will also suggest us some of the ways to improve the performance of the retailers to increase their total revenue collection.

3 Introduction

The data which we are analysing in this project is from IBM Website. This data set contains **88475 rows**, **11 Data Columns** of almost **144 products**. The dataset is from retailers of **21 different countries**. The data is organized is a way that the first columns contains the name of the country and then the next is the name of the retailer and its selling method type. Among all the other features the last one is the revenue generated which we will be focusing more on. The focus area of our problem is to identify the reasons why certain countries are generating more revenue than others.

4 Implementation

The very first step of the project was to make sense out of the raw data. We created different pivot tables out of the dataset to group the data logically. The below figure depicts the dashboard of the interactive web Application. It is the homepage of the web Application which shows overview of the entire dataset stating all the 21 countries along with their corresponding total revenue. The first entry in the table is the country with the highest revenue and the last is the least revenue generating country. The 3D pie chart explains the country-wise percentage contribution to the total revenue. The data table on the right shows the statistics regarding total revenue which also provides the sorting, pagination and filtering capability. The same statistics can be visualize when user hovers over the pie chart.



21
Countries

144
Products

Country-wise Revenue

Sales Revenue Data

Show 7 entries Search:

	Country	Revenue
1	United States	650810960.52
2	Japan	281665498.62
3	China	248823216.53
4	Canada	246887664.87
5	France	219523642.18
6	United Kingdom	219223087.62
7	Germany	200773663.38

Showing 1 to 7 of 21 entries Previous 1 2 3 Next

From the above analysis the least performing countries can be deduced. We have to find the reason for the poor performance of the countries. The next step is to analyse the retailer wise revenue shares for the least performing countries. The figure below displays the retailer wise revenue shares, which is in the Sales Menu, Retailer-wise revenue tab.

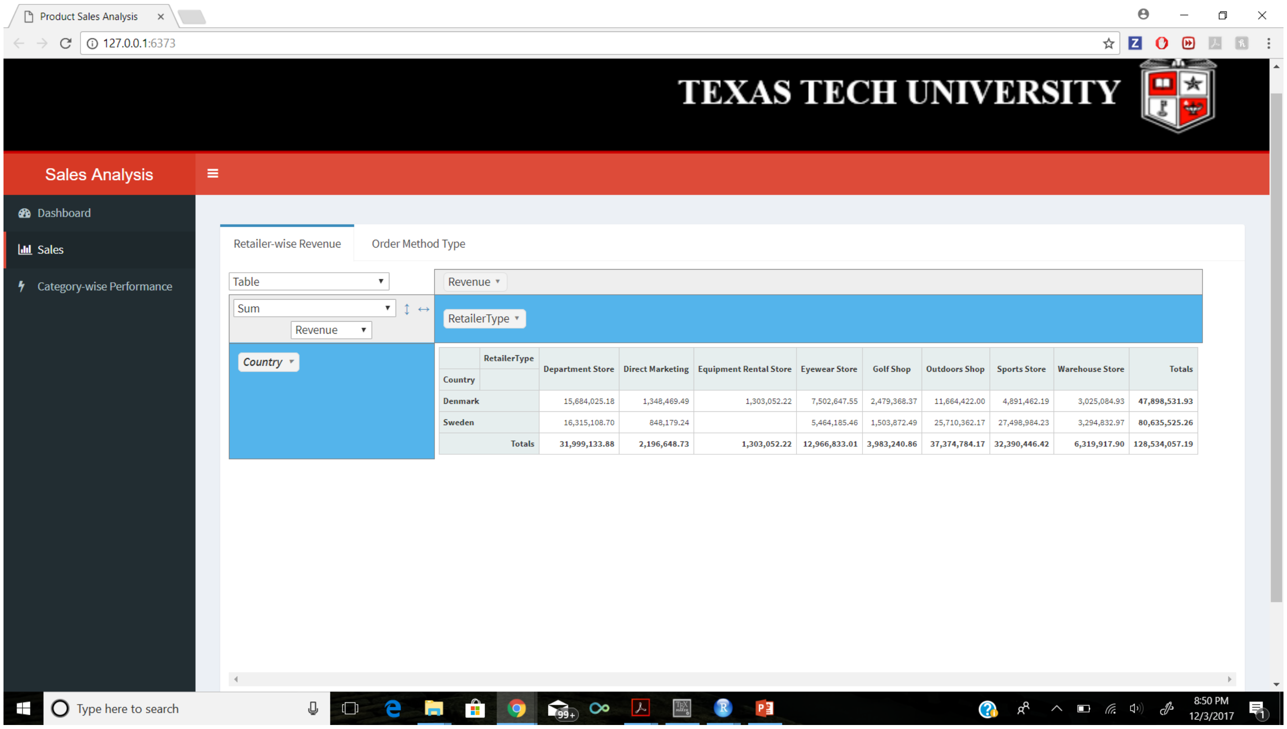
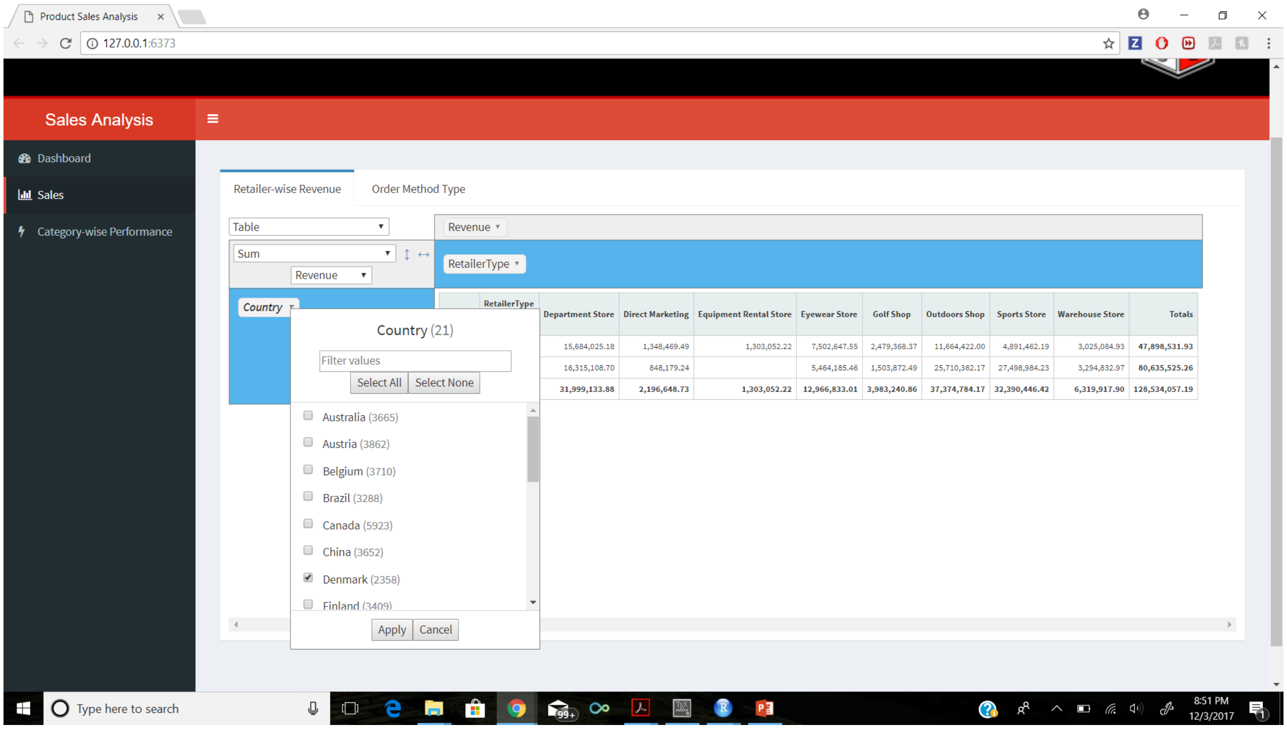


Table
Sum
Revenue
Revenue
RetailerType
Country

	RetailerType	Department Store	Direct Marketing	Equipment Rental Store	Eyewear Store	Golf Shop	Outdoors Shop	Sports Store	Warehouse Store	Totals
Country										
Denmark		15,684,025.18	1,348,469.49	1,303,052.22	7,502,647.55	2,479,368.37	11,664,422.00	4,891,462.19	3,025,084.93	**47,898,531.93**
Sweden		16,315,108.70	848,179.24		5,464,185.46	1,503,872.49	25,710,362.17	27,498,984.23	3,294,832.97	**80,635,525.26**
	Totals	**31,999,133.88**	**2,196,648.73**	**1,303,052.22**	**12,966,833.01**	**3,983,240.86**	**37,374,784.17**	**32,390,446.42**	**6,319,917.90**	**128,534,057.19**

The least two performing countries data is displayed by default, if we want to view data for other countries, we have to select the countries from drop down filter as seen in the figure below.



Country (21)

Filter values

Select All | Select None

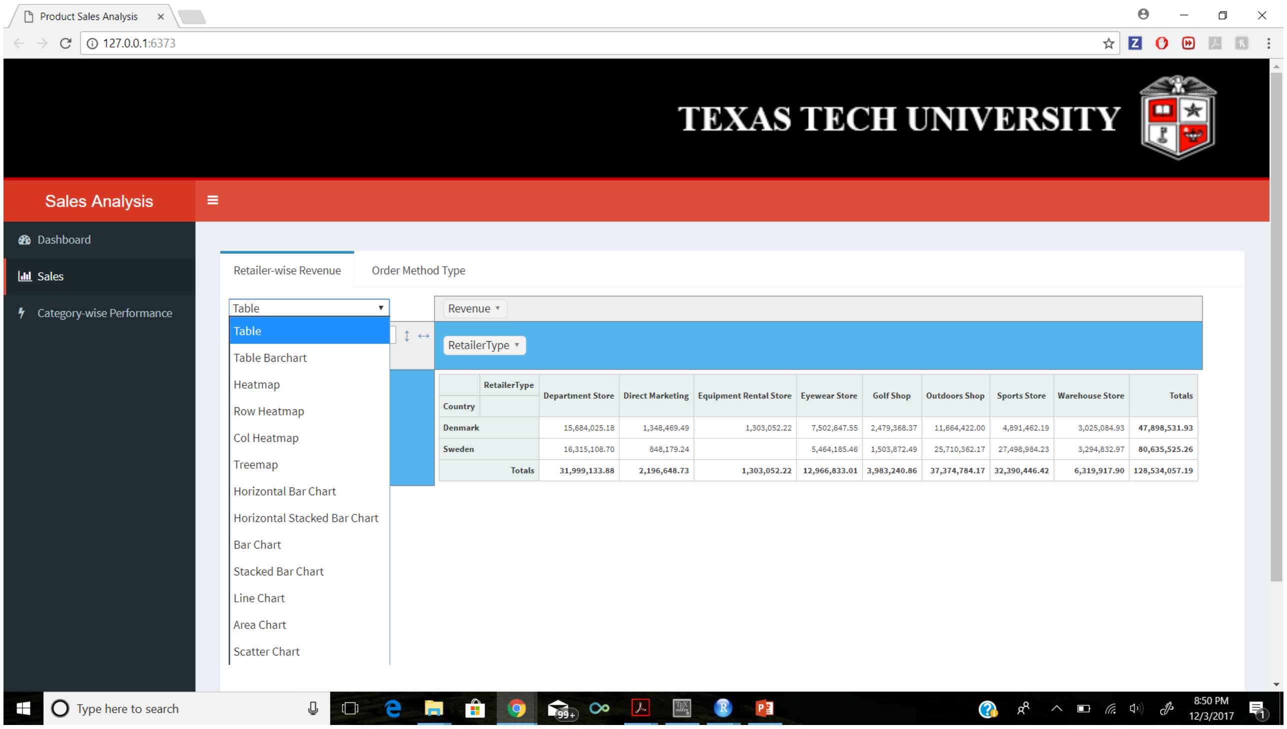
- □ Australia (3665)
- □ Austria (3862)
- □ Belgium (3710)
- □ Brazil (3288)
- □ Canada (5923)
- □ China (3652)
- ☑ Denmark (2358)
- □ Finland (3409)

Apply | Cancel

RetailerType	Department Store	Direct Marketing	Equipment Rental Store	Eyewear Store	Golf Shop	Outdoors Shop	Sports Store	Warehouse Store	Totals
	15,684,025.18	1,348,469.49	1,303,052.22	7,502,647.55	2,479,368.37	11,664,422.00	4,891,462.19	3,025,084.93	**47,898,531.93**
	16,315,108.70	848,179.24		5,464,185.46	1,503,872.49	25,710,362.17	27,498,984.23	3,294,832.97	**80,635,525.26**
	31,999,133.88	**2,196,648.73**	**1,303,052.22**	**12,966,833.01**	**3,983,240.86**	**37,374,784.17**	**32,390,446.42**	**6,319,917.90**	**128,534,057.19**

Additionally, we can view different types of graphs. Select from the first drop down with default value Table. The option available are
1. Pie charts
2. heat map
3. bar chart
4. Area charts
5. Tree map
6. Line chart etc.

Screen-shot depicts the feature.



	RetailerType	Department Store	Direct Marketing	Equipment Rental Store	Eyewear Store	Golf Shop	Outdoors Shop	Sports Store	Warehouse Store	Totals
Country										
Denmark		15,684,025.18	1,348,469.49	1,303,052.22	7,502,647.55	2,479,368.37	11,664,422.00	4,891,462.19	3,025,084.93	**47,898,531.93**
Sweden		16,315,108.70	848,179.24		5,464,185.46	1,503,872.49	25,710,362.17	27,498,984.23	3,294,832.97	**80,635,525.26**
Totals		**31,999,133.88**	**2,196,648.73**	**1,303,052.22**	**12,966,833.01**	**3,983,240.86**	**37,374,784.17**	**32,390,446.42**	**6,319,917.90**	**128,534,057.19**

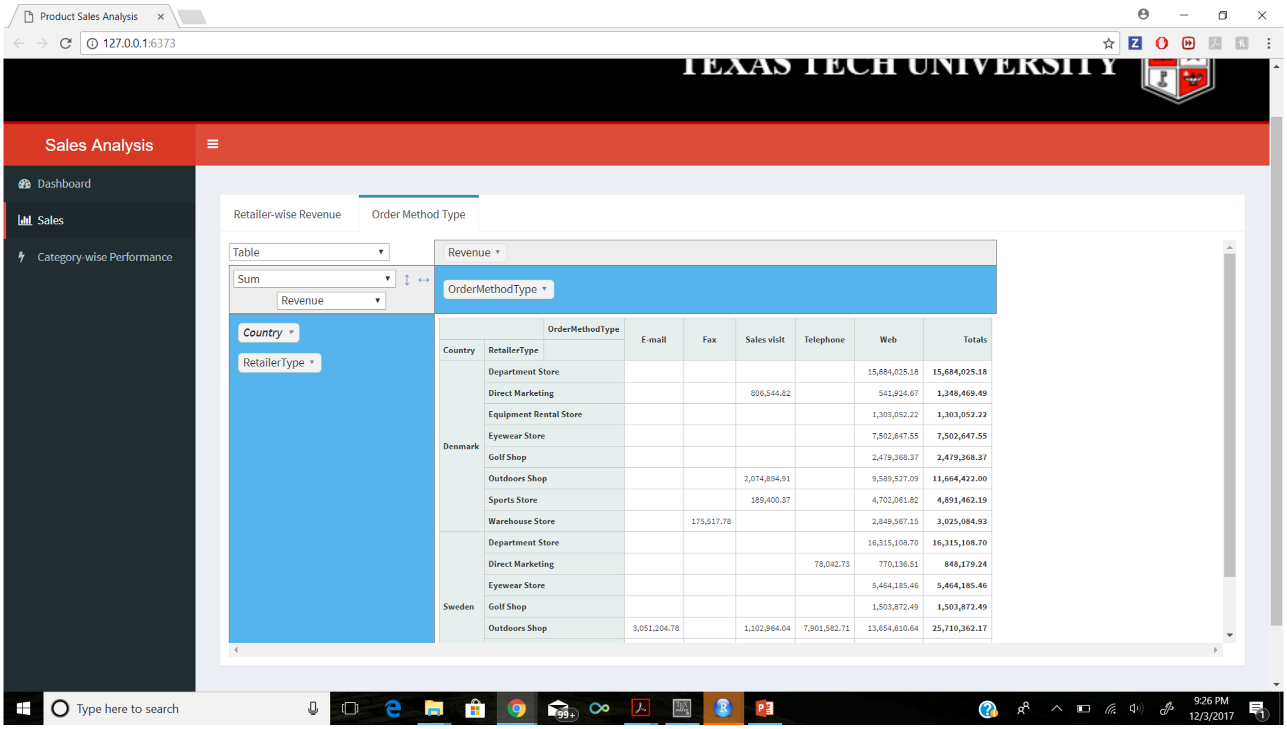
We have till yet obtained:

1. Least revenue generating countries.
2. Least performing retailers for those countries.

From the data it is observed that, the retailers have different order method types:

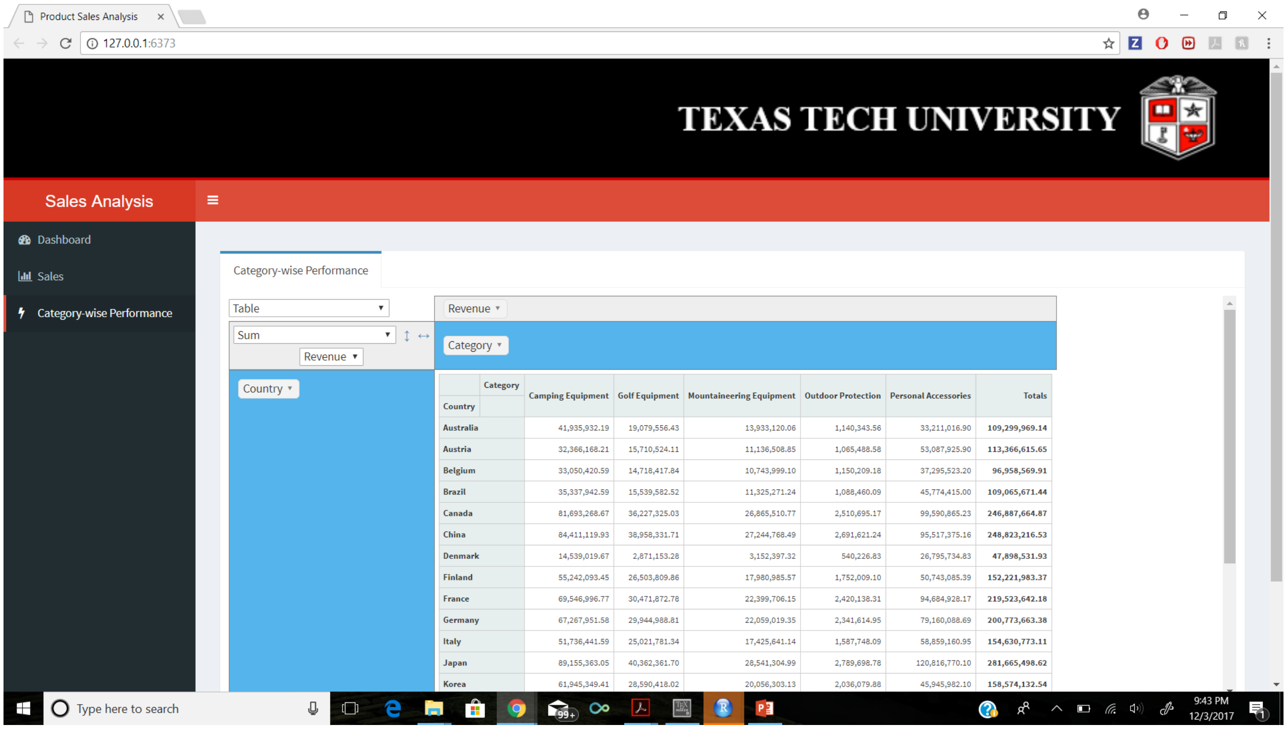
1. Web
2. Telephone
3. Fax
4. Email
5. Sales Visit etc.

Now we partition the data further, based on order method types for the above data. We analyse which method type is not used up to the mark and requires improvement. We can view the same in Sales menu Order Method Type tab. Below screen-shot depicts the same.



	OrderMethodType	E-mail	Fax	Sales visit	Telephone	Web	Totals
Country	RetailerType						
Denmark	Department Store					15,684,025.18	**15,684,025.18**
	Direct Marketing			806,544.82		541,924.67	**1,348,469.49**
	Equipment Rental Store					1,303,052.22	**1,303,052.22**
	Eyewear Store					7,502,647.55	**7,502,647.55**
	Golf Shop					2,479,368.37	**2,479,368.37**
	Outdoors Shop			2,074,894.91		9,589,527.09	**11,664,422.00**
	Sports Store			189,400.37		4,702,061.82	**4,891,462.19**
	Warehouse Store		175,517.78			2,849,567.15	**3,025,084.93**
Sweden	Department Store					16,315,108.70	**16,315,108.70**
	Direct Marketing				78,042.73	770,136.51	**848,179.24**
	Eyewear Store					5,464,185.46	**5,464,185.46**
	Golf Shop					1,503,872.49	**1,503,872.49**
	Outdoors Shop	3,051,204.78		1,102,964.04	7,901,582.71	13,654,610.64	**25,710,362.17**

We also provided an overview of category-wise performance for different countries. It can be viewed from Category-wise Performance menu.



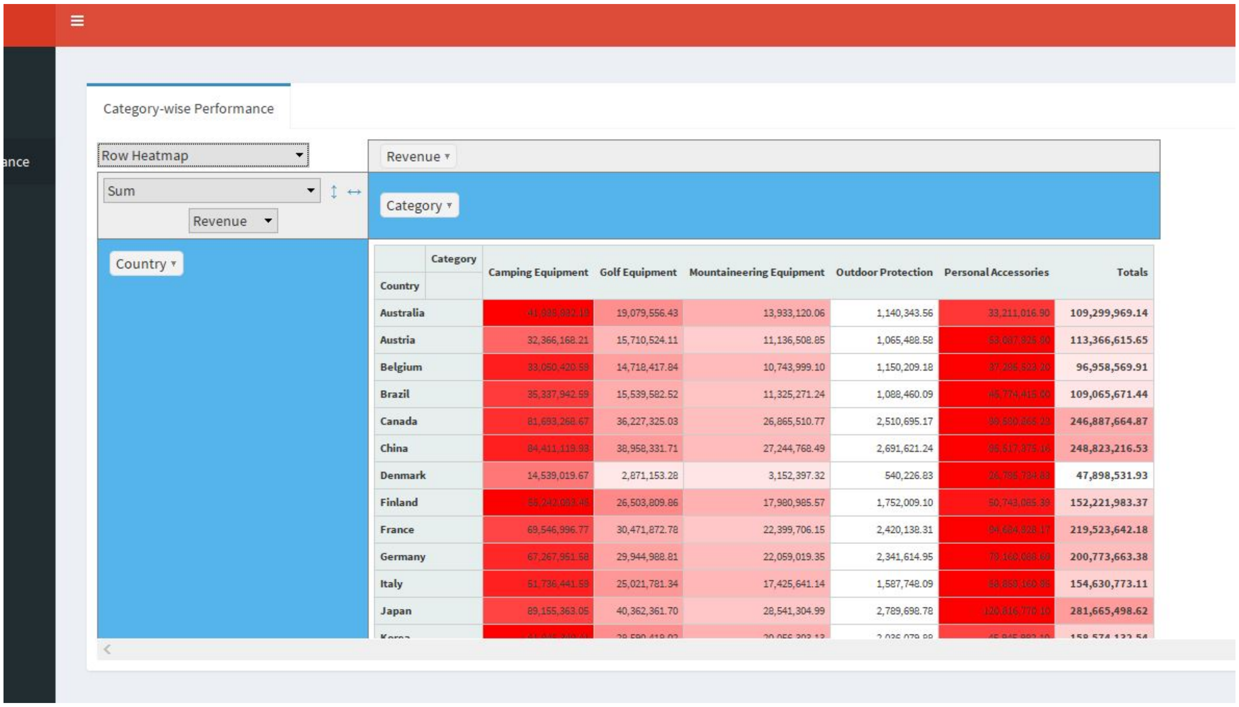
	Category	Camping Equipment	Golf Equipment	Mountaineering Equipment	Outdoor Protection	Personal Accessories	Totals
Country							
Australia		41,935,932.19	19,079,556.43	13,933,120.06	1,140,343.56	33,211,016.90	**109,299,969.14**
Austria		32,366,168.21	15,710,524.11	11,136,508.85	1,065,488.58	53,087,925.90	**113,366,615.65**
Belgium		33,050,420.59	14,718,417.84	10,743,999.10	1,150,209.18	37,295,523.20	**96,958,569.91**
Brazil		35,337,942.59	15,539,582.52	11,325,271.24	1,088,460.09	45,774,415.00	**109,065,671.44**
Canada		81,693,268.67	36,227,325.03	26,865,510.77	2,510,695.17	99,590,865.23	**246,887,664.87**
China		84,411,119.93	38,958,331.71	27,244,768.49	2,691,621.24	95,517,375.16	**248,823,216.53**
Denmark		14,539,019.67	2,871,153.28	3,152,397.32	540,226.83	26,795,734.83	**47,898,531.93**
Finland		55,242,093.45	26,503,809.86	17,980,985.57	1,752,009.10	50,743,085.39	**152,221,983.37**
France		69,546,996.77	30,471,872.78	22,399,706.15	2,420,138.31	94,684,928.17	**219,523,642.18**
Germany		67,267,951.58	29,944,988.81	22,059,019.35	2,341,614.95	79,160,088.69	**200,773,663.38**
Italy		51,736,441.59	25,021,781.34	17,425,641.14	1,587,748.09	58,859,160.95	**154,630,773.11**
Japan		89,155,363.05	40,362,361.70	28,541,304.99	2,789,698.78	120,816,770.10	**281,665,498.62**
Korea		61,945,349.41	28,590,418.02	20,056,303.13	2,036,079.88	45,945,982.10	**158,574,132.54**

5 Modelling

The next step in the project was to make a multiple linear regression model. The first thing to know here is what is linear regression.

In statistics, linear regression is a linear approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case where we use one explanatory variable is called simple linear regression[2]. And the case which we will be doing here is the one with multiple dependent variables and is known as Multiple Linear Regression.

As per the data visualization, its clear that the "Outdoor Protection" product line is generating the least revenue. HeatMap below depicts the same results.



Country \ Category	Camping Equipment	Golf Equipment	Mountaineering Equipment	Outdoor Protection	Personal Accessories	Totals
Australia	41,935,932.19	19,079,556.43	13,933,120.06	1,140,343.56	33,211,016.90	109,299,969.14
Austria	32,366,168.21	15,710,524.11	11,136,508.85	1,065,488.58	53,087,925.90	113,366,615.65
Belgium	33,050,420.58	14,718,417.84	10,743,999.10	1,150,209.18	37,295,523.20	96,958,569.91
Brazil	35,337,942.59	15,539,582.52	11,325,271.24	1,088,460.09	45,774,415.00	109,065,671.44
Canada	81,693,268.67	36,227,325.03	26,865,510.77	2,510,695.17	99,590,865.23	246,887,664.87
China	84,411,119.93	38,958,331.71	27,244,768.49	2,691,621.24	95,517,375.16	248,823,216.53
Denmark	14,539,019.67	2,871,153.28	3,152,397.32	540,226.83	26,795,734.83	47,898,531.93
Finland	55,242,093.46	26,503,809.86	17,980,985.57	1,752,009.10	50,743,085.39	152,221,983.37
France	69,546,996.77	30,471,872.78	22,399,706.15	2,420,138.31	94,684,928.17	219,523,642.18
Germany	67,267,951.58	29,944,988.81	22,059,019.35	2,341,614.95	79,160,088.69	200,773,663.38
Italy	51,736,441.58	25,021,781.34	17,425,641.14	1,587,748.09	58,859,160.96	154,630,773.11
Japan	89,155,363.05	40,362,361.70	28,541,304.99	2,789,698.78	120,816,770.10	281,665,498.62
Korea	[illegible]	28,590,418.02	20,056,303.13	2,036,079.88	[illegible]	158,574,133.54

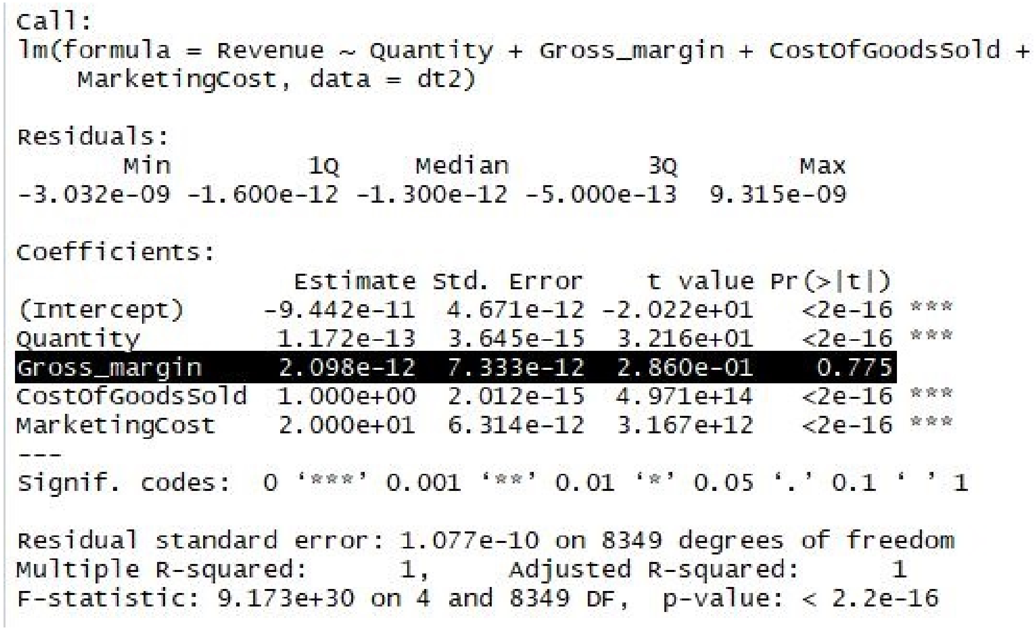
The focus area of our model targets the data for "Outdoor Protection" product line. The features used for building the model is as below which are independent variables. 1. Quantity
2. Gross margin
3. Cost of Goods Sold
4. Marketing Cost

Below is a code snippet for the lm() function we used for regression modelling.

```
lm.mod <- lm(Revenue ~ Quantity + Gross_margin + CostOfGoodsSold + MarketingCost ,
    data = dt2)
summary(lm.mod)
```

5.1 Results

5.1.1 First Model



```
Call:
lm(formula = Revenue ~ Quantity + Gross_margin + CostOfGoodsSold +
    MarketingCost, data = dt2)

Residuals:
       Min         1Q     Median         3Q        Max
-3.032e-09 -1.600e-12 -1.300e-12 -5.000e-13  9.315e-09

Coefficients:
                  Estimate Std. Error    t value Pr(>|t|)
(Intercept)     -9.442e-11  4.671e-12 -2.022e+01   <2e-16 ***
Quantity         1.172e-13  3.645e-15  3.216e+01   <2e-16 ***
Gross_margin     2.098e-12  7.333e-12  2.860e-01    0.775
CostOfGoodsSold  1.000e+00  2.012e-15  4.971e+14   <2e-16 ***
MarketingCost    2.000e+01  6.314e-12  3.167e+12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.077e-10 on 8349 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 9.173e+30 on 4 and 8349 DF,  p-value: < 2.2e-16
```

Before digging deep into the results, knowing the components of the result is a better idea.

The first component of the regression model here is the **Residuals:** The residuals are the difference between the actual values of the variable we are predicting and predicted values from the regression (y -). If our residuals are normally distributed, this indicates the mean of the difference between our predictions and the actual values is close to 0 which is good[1].

Estimated Coefficients: The estimated coefficient is the value of slope calculated by the regression. It might seem a little confusing that the Intercept also has a value, but just think of it as a slope that is always multiplied by 1. This number will obviously vary based on the magnitude of the variable we are inputting into the regression, but it's always good to spot check this number to make sure it seems reasonable[1].

Standard Error: We are getting an estimate here in the results. So, obviously there would be some error associated with it. It is the measure of the variability in the estimate for the coefficient.

T- value: Score that measures whether or not the coefficient for this variable is meaningful for the model. We probably won't use this value itself, but for the knowledge it is used to calculate the p-value and the significance levels.

P- value: Probability that the variable is NOT relevant. Everyone wants this number to be as small as possible. If the number is really small, R will display it in scientific notation. In or example 2e-16 in our case means that the odds that parent is meaningless is about 15000000000000000[1].

Significance Legends: The more punctuation there is next to your variables, the better.

Blank=bad, Dots=pretty good, Stars=good, More Stars=very good.
Here in our case three of the variables are very significant for the model.

Multiple R-Squared: The r square values is a metric to evaluate goodness of fit of the model. with 1 being the highest and o being the worst fit.

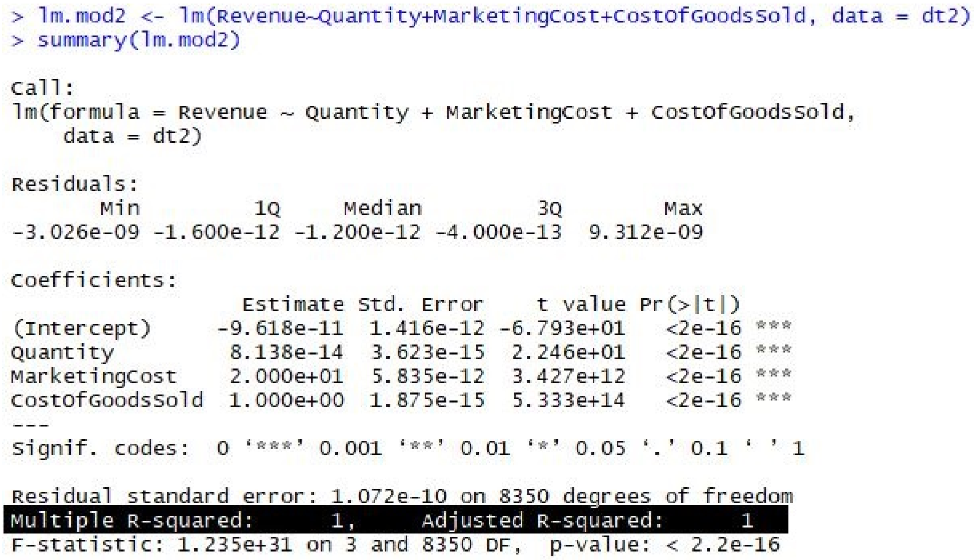
Understanding the Result: As described above the significance legends shows that how significant is the independent variable is for the model. The summary object of our regression model clearly shows that Gross margin is least significant for the model to predict the Revenue shares. Hence it was removed from the model.

5.1.2 Second Model

Code snippet to make the model again after removing the gross margin.

```r
lm.mod <- lm(Revenue ~ Quantity + CostOfGoodsSold + MarketingCost , data = dt2)
summary(lm.mod)
```

The results for second model is as below.



```
> lm.mod2 <- lm(Revenue~Quantity+MarketingCost+CostOfGoodsSold, data = dt2)
> summary(lm.mod2)

Call:
lm(formula = Revenue ~ Quantity + MarketingCost + CostOfGoodsSold,
    data = dt2)

Residuals:
       Min         1Q     Median         3Q        Max
-3.026e-09 -1.600e-12 -1.200e-12 -4.000e-13  9.312e-09

Coefficients:
                  Estimate Std. Error    t value Pr(>|t|)
(Intercept)     -9.618e-11  1.416e-12 -6.793e+01   <2e-16 ***
Quantity         8.138e-14  3.623e-15  2.246e+01   <2e-16 ***
MarketingCost    2.000e+01  5.835e-12  3.427e+12   <2e-16 ***
CostOfGoodsSold  1.000e+00  1.875e-15  5.333e+14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072e-10 on 8350 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.235e+31 on 3 and 8350 DF,  p-value: < 2.2e-16
```

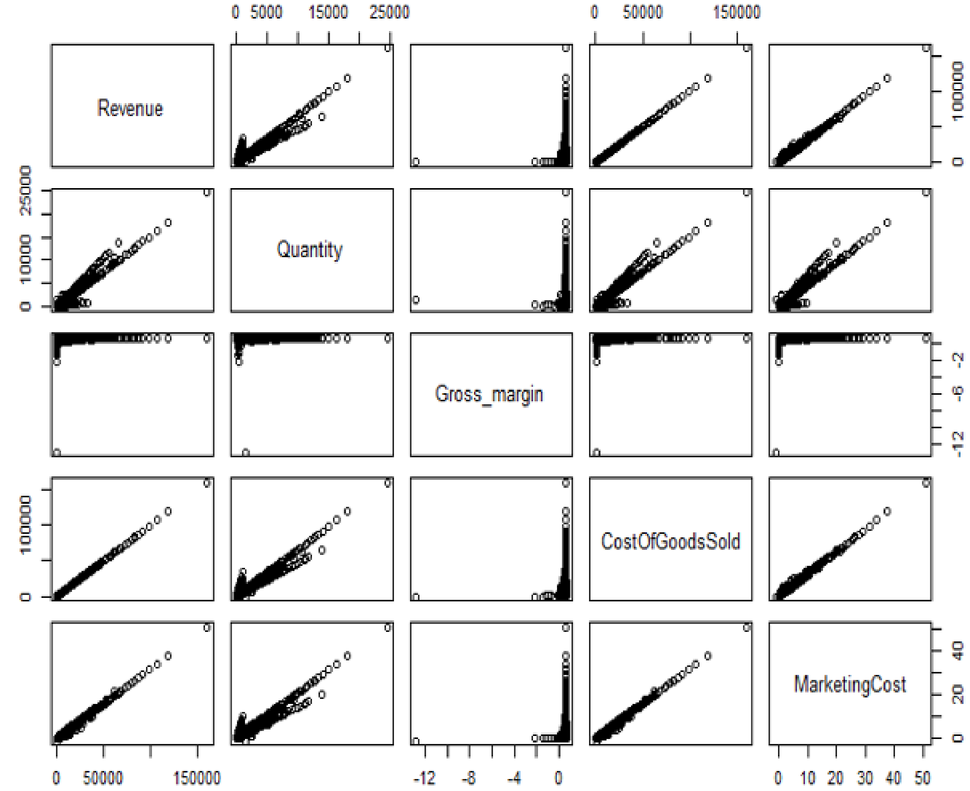
The results shows that all the variables are highly significant to the the model.

However, we noticed that multiple R-squared value of the model is 1, which is the highest value a model can have and also it is very unlikely. Hence, we analysed further to figure out the reason about the perfect fit, thus increasing the scope of our project.

5.2 Extended Analysis

We plotted the Scatter plot Matrix to visualize if there is any multicollinearity between the independent variables used in the modelling. The Scatter Plot Matrix shows the results below.

5.2.1 Scatter Plot Matrix

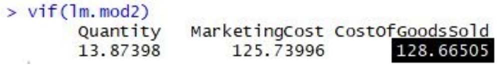


We can clearly see from the above result that the independent variables marketing cost and Cost of Goods Sold have a high degree of collinearity. But to strengthen our decision about the multicollinearity we performed VIF which stands for Variance Inflation Factor.

5.2.2 Variance Inflation factor

In statistics, the variance inflation factor (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity[2]. The "VIF" function is available in the "car" R package. Analysis of the magnitude of multicollinearity is done by considering the size of the VIF. A rule of thumb is that if VIF is greater than 10 then multicollinearity is high, which is not good.

Result:



```
> vif(lm.mod2)
      Quantity   MarketingCost CostOfGoodsSold
      13.87398       125.73996       128.66505
```

The result shows that there is high multicollinearity among the independent variables, which means if one of the variables is removed from the model it will not affect the goodness of the model much. So, to decrease the multicollinearity from model the key idea is to remove the variable with the highest VIF. Here as we highlighted in the results, Cost of Goods Sold have the highest Variance

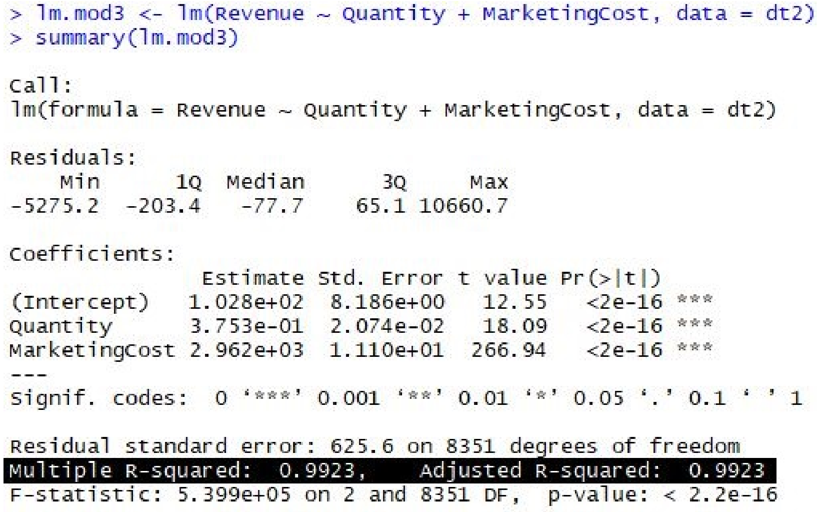
Inflation Factor.

We removed the variable and then created another regression model with only two independent variables. Code snippet is given below.

```
lm.mod <- lm(Revenue ~ Quantity + MarketingCost, data = dt2)
summary(lm.mod)
```

5.2.3 Third Model

The result from the third model is given below:

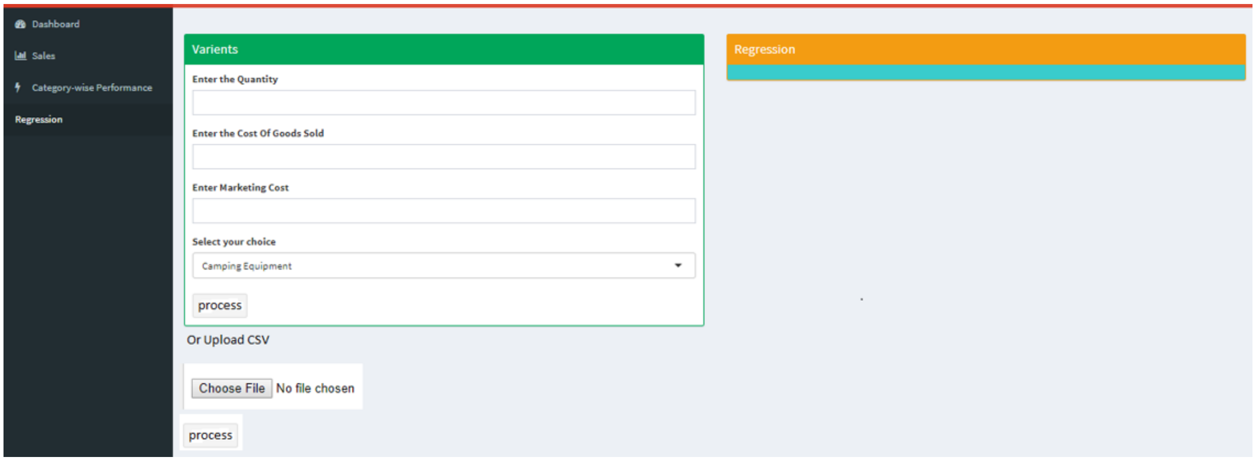


```
> lm.mod3 <- lm(Revenue ~ Quantity + MarketingCost, data = dt2)
> summary(lm.mod3)

Call:
lm(formula = Revenue ~ Quantity + MarketingCost, data = dt2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5275.2  -203.4   -77.7    65.1 10660.7 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.028e+02  8.186e+00   12.55   <2e-16 ***
Quantity      3.753e-01  2.074e-02   18.09   <2e-16 ***
MarketingCost 2.962e+03  1.110e+01  266.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 625.6 on 8351 degrees of freedom
Multiple R-squared:  0.9923,	Adjusted R-squared:  0.9923 
F-statistic: 5.399e+05 on 2 and 8351 DF,  p-value: < 2.2e-16
```

Notice the highlighted part in the result. The multiple R-squared value changed from 1 to 0.9. That is a good sign.

6 Integration of model with web application

We integrated the model obtained with the screen. Where user can simply throw in the data and get the predicted revenue generated for the fed in data. Here is a glimpse of that.



- The first column here is to fill in the quantity.

- The second column is to fill the Cost of Goods Sold.
- The third column is to fill Marketing cost.
- The fourth one is to select the category.

7 Conclusions

Summary for the analysis:

- Top countries for the most revenue shares are:
  - United States 17%
  - Japan, China, Canada 7%
- Bottom 2 countries for least revenue shares are:
  - Sweden 2%
  - Denmark 1%
- In Denmark, the poor performers are:
  - Equipment Rental Store
  - Direct Marketing Team
- In Sweden, the poor performers are:
  - Direct Marketing Team
  - Golf Shop
- Reasons:
  - Direct Marketing Team , Equipment Rental Store and Golf Shop are not utilizing all of the resources, that's why they are the least performer.
  - Outdoor protection is the category which is not doing well globally.
  - Moreover Department Store, Eyewear Store, Warehouse Store and Sports Store are also not utilizing resource to the fullest.
- Possible cause:
  - lack of knowledge.

> And the problem of this lack of knowledge can be rectified by using the web application we created to predict the better revenue by analyzing about the utilization of the resources.

8 Citations

1. http://blog.yhat.com/posts/r-lm-summary.html
2. Wikipedia